Published in final edited form as:
Neuroimage. 2016 January 01; 124(Pt B): 1065–1068. doi:10.1016/j.neuroimage.2015.10.079.

Sharing the wealth: Neuroimaging data repositories

Simon Eickhoff,
Institute of Clinical Neuroscience and Medical Psychology, Heinrich-Heine University Düsseldorf and Institute of Neuroscience and Medicine (INM-1), Research Center Jülich, Germany

Thomas E. Nichols,
Department of Statistics & WMG, University of Warwick, Coventry, UK

John D. Van Horn, and
USC Mark and Mary Stevens Neuroimaging and Informatics Institute, University of Southern California, Los Angeles, CA, USA

Jessica A. Turner
Psychology Department & Neuroscience Institute, Georgia State University, Atlanta, GA, USA, Mind Research Network, Albuquerque, NM, USA

Neuroimaging by means of functional and structural magnetic resonance imaging (MRI) has become the by far most widely used tool to study human brain organization in vivo. Since the 1990s, applications of MRI in neuroimaging have been almost exclusively devoted to the assessment of task-evoked changes in neuronal activity in the context of cognitive neuroscience experiments. These studies usually involved no more than a dozen subjects, represented a logical extension of experimental psychology, with differential brain responses to different tasks or stimuli serving as markers for differential processing modules. Over the last decade, however, the landscape has changed dramatically through three closely related developments. The first was the rapid advancements in high-resolution structural neuroimaging, i.e., methods and approaches that allowed studying detailed aspects of brain structure, function and connectivity outside of the context of experimental task paradigms. Using voxel-based morphometry, cortical thickness analysis or related techniques, morphological information derived from T1-weighted anatomical scans may be computed, compared across groups or in relation to phenotypic variability. Methods for the analysis of diffusion-weighted imaging (DWI) have also been the focus of considerable development allowing the characterization of fiber pathways in the brain and the detailed quantification of white-matter architecture. Finally, resting-state functional MRI (rs-fMRI) has provided many new and exciting avenues for exploring the brain's functional connectivity and network architecture.

The efficiency and growing prominence of various neuroimaging techniques has served as a catalyst for the other two major trends in the field. One is the steep increase in the number of subjects that are included in any given study. While the earliest neuroimaging studies used as few as 4–6 subjects, and a decade ago studies with more than 12–16 participants were still unusual for task-based fMRI studies, it was quickly realized that task-free approaches allowed the collection and analysis of much larger samples. These investigations hence led the way into the assessment of dozens to hundreds of subjects, which is commonplace today.

The surge in study sample sizes, providing superior statistical power and reduced vulnerability to spurious effects, combined with the potential to easily reuse task-free data across many different analyses (since the obtained parameters are not task dependent and hence don't require detailed descriptors of experimental procedures) has laid the groundwork for a growth in data sharing. Consequently, the data from human neuroimaging studies, both with and without cognitive tasks, are now being more widely shared, pooled, and re-analyzed.

Given the logistic and monetary expenses associated with MRI scanning, the effort required to recruit subjects, and assess them and, of course, the time invested by the participants themselves, it may be argued that sharing data and allowing re-analysis in different contexts is almost an inevitable prerequisite for further advancement in the field of neuroimaging. It moreover resonates well with the increasingly recognized need for transparent and reproducible science.

All of this motivates the present special issue. This issue provides an admittedly incomplete snapshot of the state of neuroimaging data sharing in 2015, with more than 40 repositories or datasets being introduced and described in the first issue, and a second issue to follow. The vast amount of data that is available to the community through these repositories attests to the rapid growth of data sharing, and the transformation of neuroimaging into a data-intense field aimed at the understanding of fundamental principles and inter-individual variability.

The repositories are reported on here briefly; the point of the issue was not to tout particular findings or study nuances of database structures, but to make each repository clearly public to the research community. These are short reports intended to explain what data repository has, how one gains access to it, whether it is a static archive or a growing repository, and other basic facts. Upon reviewing the different neuroimaging data resources described in this issue, we noted several major trends but also highly divergent factors among them. We comment on several of the major topics we identified to outline the field of neuroimaging data-sharing as a whole, but also to point to important further developments. The full list of papers and resources for this issue is listed in Table 1.

One important distinction between the different initiatives summarized in this issue relates to the scope of the data-sharing. It was initially expected that most if not all submissions would cover both an infrastructure for sharing as well as the specific datasets available through it; and many of the repositories are serving up raw data from extremely large studies or very small ones. Examples of the large datasets are the Pediatric Imaging, Neurocognition and Genetics (PING) (Jernigan et al., this issue), a single study of 1493 children, or the Philadelphia Neurodevelopmental Cohort (Satterthwaite et al., this issue), a single study of about 1000 datasets; while smaller ones can be found in the Brainomics system, for example, with a single static dataset of 94 subjects with multiple imaging modalities and genetic data (Orfanos et al., forthcoming). The first and largest sections of papers are repositories and datasets shared through various repositories (e.g., NITRC (Kennedy et al., this issue; Karayanidis et al., this issue; and Jernigan et al., this issue) or XNAT (Herrick et al., this issue; Harrigan et al., this issue; and Alpert et al., this issue)). A new addition in this data sharing issue is mediated data sources, such as SchizConnect and GAAIN (Wang et al. this issue; Neu et al., this issue), which serve up data from other existing databases in a unified framework, making it easier to find data across existing resources.

But neuroimaging data sharing has developed into a broad field in the past decade, and has significant heterogeneity in scope. Most fundamentally, several resources provide aggregate data such as statistical maps from group analyses (Gorgolewski et al., this issue; Brown et al., this issue), or results from neuroimaging meta-analyses (Reid et al., this issue). In contrast to resources providing raw data of individual subjects (human or animal (Sawiak et al., this issue)), these analysis repositories obviously represent a high level of abstraction and will hence primarily be used as a priori information for the analysis or interpretation of future neuroimaging studies. The respective data sharing projects hence resemble brain atlases more than resources for neuroimaging data (e.g., Sawiak et al.). With that variety of resources available, one of the major challenges that remain is to integrate these resources into a comprehensive human brain atlas of structure, function and connectivity.

However, there is also a considerable spectrum of concepts among repositories that provide raw data. At one end of the spectrum, single or collaborative labs are sharing a particular dataset they have acquired, sometimes through data sharing national resources such as NDAR or dbGAP, e.g., (Pierpaoli et al., this issue; Jernigan et al., this issue). At the other end of the spectrum, there are flexible data management and sharing repositories that allow the deposition of various kinds of data by any user (Kennedy et al., this issue). In the former case, the actual data being shared defines the database, and the archive is static (e.g., Keator et al., this issue) and the addition of new or different data by the users of that resource is usually not feasible. In contrast, repositories may initially contain data from the hosting institution, e.g. managing their scanning center output, but is meant to be populated by any user (e.g. Landis et al.; Herrick et al.; Harrigan et al.; Book et al; and Alpert et al., this issue). As a special case of the latter, this special issue describes several repositories that were primarily intended to store, manage and share data internally but have been opened to the scientific community (see e.g., Landis et al., Harrigan et al., Crawford et al., Herrick et al.). It may be noted, though, that the distinction between a single shared dataset and a repository to be populated does not necessarily imply different strategies for database management. While in some cases, sharing of a particular dataset may be accomplished with rather little infrastructure beyond a webserver, this can be substantially more complex in the case of larger, prospective data-sharing initiatives such as Human Connectome Project (e.g., Hodge et al., and Fan et al., this issue).

While some databases are very broad and have an infrastructure that can contain many different types of data, it quickly becomes clear that many databases are a reflection of the internal differentiation within the field of neuroimaging. In particular, the papers in this issue reflect two distinct types of focus that can be found in existing databases. Either these are specific to a particular imaging modality, e.g. arterial spin labeling or MEG (Shin et al., Tardif et al., Niso et al.), or they target a particular population. Examples of the latter would be databases covering particular patient populations or developmental aspects (Satterthwaite et al.; Vaillancourt et al; Karayanidis et al; Jernigan et al; and Seghier et al.). While this fragmentation evidently reflects the different interests of researchers in the field of imaging neuroscience, it impedes the integration of data across datasets. The presence of amore sophisticated and comprehensive database structure of course creates the possibilities of accessing the different databases through Application Program Interfaces (APIs) and the possibility of automated data extraction and federated queries. As reflected in the current special issue, we note that such capabilities are still far from commonplace and, if present, mainly are found in repositories rather than individually shared datasets. See Wang et al. or the SchizConnect example, and Neu et al. for the GAAIN example, leveraging several existing databases of very different design to create federated queries, so that a single query will retrieve hundreds of results across multiple repositories. The extension of automated queries and federated search capabilities is essential, e.g., for joint analyses of healthy control subjects obtained for different disorders for methods development, the investigation of neurobiological variance in the general population or the assessment of medication effects.

The combination of shared data would without doubt open completely new perspectives in the robust investigation of brain organization and inter-individual variability, by pooling across many thousands of subjects from different populations, scanned on different machines using different imaging sequences. Any finding emerging from such large-scale integration could hence be considered representative knowledge about the human brain that is unlikely to be driven by any particular technical or biological confounder.

Reflecting on this possibility, however, raises another issue that becomes painfully obvious when looking at the state of neuroimaging data-sharing in 2015: the lack of consistency in the provided information. On the neuroimaging side, almost all MRI-based resources (which represent the majority of the available data sharing initiatives) provide structural T1 weighted imaging and most also make BOLD resting-state scans available (albeit using vastly heterogeneous protocols). In turn, additional images, such as diffusion-weighted sequences, are considerably more rare. The most striking heterogeneity, however, relates to the phenotypical (behavioral) data that is being shared. This not only pertains to the amount of information that is released, which varies widely between databases and samples, but also to the actual measures that were obtained and made available. Several trends may be observed. In general, datasets shared by an individual site through a dedicated portal usually contain substantially more phenotypical information than data that is available through repositories that are populated by the users. Furthermore, as an important obstacle to cross-dataset analyses, different instruments are often used to assess a particular aspect such as cognitive performance, personality or psychopathology. Finally, it is noteworthy that some aspects are mostly consistently reported in some settings but not in others; e.g., ethnicity is usually reported in datasets from the US but rarely in those from Europe. Consequently, analyses across data provided by different databases and repositories are often restricted to assessing relationships to the most basic phenotypic measures, i.e., age and gender.

This combining of data across studies and sources raises the important issue of quality control, which in the case of neuroimaging databases, encompasses two distinct aspects. The first is quality control in the more narrow sense, i.e., ensuring that the data is of sufficient quality to be usable. Quality control (QC) in that sense includes checking for the presence of artifacts, quantifying signal-to-noise ratios, potentially identifying outliers and related issues. Several of the repositories include QC measures or pipelines to compute them, but it is often not included with the data. While potentially useful, there are few generally accepted standards and different use cases may require different stringency levels on the rejection of "bad data". It seems appropriate to delegate the responsibility for deciding which data is usable from the technical point of view to the user; however, the ability to access the data across all the repositories listed in this issue highlights that now is the time to assess how basic QC measures varies across datasets and how they affect effect sizes across studies. The second and much more important aspect of QC, however, is correctness, i.e., whether the data that is downloaded represents what it should represent. In other words, does the neuroimaging data that can be obtained match the technical and phenotypical (patient status, age, etc.) descriptions? This crucially needs to be ensured by the data provider when uploading and the database curators when linking to it. Many but not all of the repositories have addressed this to the best extent possible, requiring double entry, reviewing data labels or incorporating electronic data capture directly; others have relied on the investigators providing the data. The papers in this issue specifically report on those procedures for each repository, so that the user can be informed, and take advantage of the details that some repositories may provide.

Surprisingly, one point has been touched upon rather rarely in the different papers but which may become crucial in the future, given the expected increase in research and hence papers based on shared data: The detection, reporting and resolution of errors. Errors naturally happen when performing research even at the highest standards and they can happen both with respect to the uploaded data as well as the database itself. Hence, there is a distinct possibility that users will in good faith download data that is actually incorrect and consequently publish factually wrong results. Two developments would be exciting to see happen soon, given the amount of research that is currently being performed on legacy data that was not acquired by the person or team doing the analyses.

First, data-providers and database-managers should strive to make sure that there is a way to notify those that downloaded and presumably used the data about any important corrections. Evidently, this is easiest for repositories that require user registration and can track who has downloaded a given dataset. Here solutions need to be established for the communication between data-providers and database curators and the relay of this information to the user community, e.g., through notification boards or mailing lists. There should be a robust way of versioning data in repositories, so authors can clearly and consistently report the use of any updated data over time. Second, the research community will eventually need to establish procedures for articles describing well-conducted research which were published after peer review, but contain objectively false results and hence conclusions due to errors in the downloaded data. The possibility of this can only increase as more data becomes available. While a retraction may seem severe, leaving these objectively incorrect findings to contribute to the current literature is probably likewise not an optimal solution. While the degree of necessary adjustments will undoubtedly vary from case to case, a standard mechanism allows the authors that were notified of database errors to publish a correction of the results and if necessary discussion as part of good scientific practice.

Finally, one aspect of data sharing that is handled very differently across the resources described in this issue is how the data can be accessed. The most common model is to allow access to the shared data following a registration and possibly a short application for usage describing what kind of research will be performed (e.g., Landis et al., and Alpert et al.). There is, however, a broad range of other solutions, ranging from direct download without any registration (e.g., Keator et al., this issue, and Poldrack et al., forthcoming), to more formal application processes (e.g. Price et al., and Knudsen et al.) and finally, databases that pose substantial hurdles for data access. The latter may include data that is only made available through an official collaboration (e.g., those described in Sethi et al. and Mazoyer et al.), repositories that require substantial institutional commitments in a Data Usage Agreement prior to data access, and databases to which access is only granted if researchers upload data themselves (e.g., Labus et al.). Intuitively, one would argue that ease of access should be an important and universally positive aspect. However, beyond the consideration of error-correction and follow-up raised above, there is the consideration of data security, anonymization and ethics, particularly with regard to imaging data when combined with genetics or in samples with rare diseases. Dealing with human subjects, whose privacy must be protected and that have consented to participating in a research project under specific terms, these aspects are of utmost importance in the context of data sharing. Consequently, completely free access to shared data is predominantly though not only found for aggregate (multi-subject) data, i.e., resources that we characterized as brain atlases or templates above, where the possibility of re-identification and loss of confidentiality has been minimized.

Summing up, neuroimaging data sharing is becoming a major aspect of our field, given that neuroimaging data is reusable and researchers have realized the need for larger samples to provide robust information on neurobiology and its variability. Estimating from the papers in this issue and the forthcoming issue, there are currently tens of thousands of imaging datasets across multiple disorders or developmental stages from hundreds of neuroimaging studies in multiple modalities, available to the research community for re-use, re-analysis, training, and novel discovery. This is a hugely positive development for cognitive neuroscience, neuropsychiatry, and neurology, and its importance cannot be understated. We encourage researchers to take full advantage of what is currently available.

Given the above considerations on the current state of the field, several issues seem key to foster the future growth of the field. The first is an increased integration between databases and (ideally) a better homogeneity of the obtained measures to allow using more of the available data at once. Second, the distribution of responsibilities related to quality control needs to be clear between data providers, database curators and importantly also the users. Third is the establishment of procedures at the level of the database, the investigator, and the subsequent journal related to how errors in the databases are communicated and research performed on that data is corrected. Finally, sensible consensus standards mediating between the protection of human subjects and the idea of open science need to be agreed on and then further developed in light of the growing technical possibilities.

Acknowledgments

JT was supported by funding from NIH (NIMH), 1 R01 MH097435-01A1 and 1R56MH097870-01. SBE was supported by the Deutsche Forschungsgemeinschaft (DFG, EI 816/4-1, LA 3071/3-1, EI 816/6-1), the National Institute of Mental Health (R01-MH074457), the Helmholtz Portfolio Theme “Supercomputing and Modeling for the Human Brain” and the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 604102 (Human Brain Project). TEN was supported by the Wellcome Trust and NIH grants U54MH091657-03, R01 EB015611-01 & R01 NS075066-01A1. JVH was supported by NIH grants U24ES026465-01, R44NS081792-03A1, P41EB015922-18, and U54EB020406-02.

Table 1

Resources represented in this issue.

Title	Author
1 The NITRC image repository	Kennedy, D.N. et al.
2 The Function Biomedical Informatics Research Network data repository	Keator, D.B. et al.
3 The image and data archive at the Laboratory of Neuro Imaging	Crawford, K. et al.
4 COINS Data Exchange: an open platform for compiling, curating, and disseminating neuroimaging data	Landis, D. et al.
5 Neuroimaging data sharing on the neuroinformatics database platform	Book, G.A. et al.
7 XNAT central: open sourcing imaging research data	Herrick, R. et al.
8 Vanderbilt University Institute of Imaging Science Center for Computational Imaging XNAT: a multimodal data archive and processing environment	Harrigan, L. et al.
9 ConnectomeDB—sharing human brain connectivity data	Hodge, R. et al.
10 MGH—USC human connectome project datasets with ultra-high b-value diffusion	MRI Fan, Q. et al.
11 The Philadelphia Neurodevelopmental Cohort: a publicly available resource for the study of normal and abnormal brain development in youth	Satterthwaite, D. et al.
12 Parkinson’s disease biomarkers program brain imaging repository	Vaillancourt, D. et al.
13 The diffusion tensor imaging (DTI) component of the NIH MRI study of normal brain development (PedsDTI)	Pierpaoli, C. et al.
14 The Northwestern University Neuroimaging Data Archive (NUNDA)	Alpert, K. et al.
15 The Age-ility Project (Phase 1): structural and functional imaging and electrophysiological data repository	Karayanidis, F. et al.
16 The Open Science CBS Neuroimaging Repository: sharing ultra-high-field magnetic resonance images of the brain	Tardif, L. et al.
17 The Pediatric Imaging, Neurocognition, and Genetics (PING) data repository	Jernigan, L. et al.
18 SchizConnect: mediating neuroimaging databases on schizophrenia and related disorders for large-scale integration	Wang, L. et al.
19 Sharing data in the global Alzheimer’s association interactive network Neu, S.C. et al. 20 Data integration: combined imaging and electrophysiology data in the cloud	Kini, G. et al.
21 OMEGA: the open MEG archive	Baillet, S. et al.
22 The MNI data-sharing and processing ecosystem	Das, S. et al.
23 Northwestern University Schizophrenia Data Sharing for SchizConnect: a longitudinal dataset for large-scale integration	Kogan, A. et al.
24 The Cerebral Blood Flow Biomedical Informatics Research Network (CBFBIRN) data repository	Shin, D. et al.
25 The PLORAS Database: a data repository for predicting language outcome and recovery after stroke	Seghier, M. et al.
26 The Center for Integrated Molecular Brain Imaging (Cimbi) database	Knudsen, M. et al.
27 Database integration of protocol-specific neurological imaging datasets	Pacurar, E. et al.
28 BIL&GIN: a neuroimaging, cognitive, behavioral, and genetic database for the study of human brain lateralization	Mazoyer, B. et al.
29 Pain and Interoception Imaging Network (PAIN): a multimodal, multisite, brain-imaging repository for chronic somatic and visceral pain disorders	Labus, J. et al.
30 Connected brains and minds — the UMCD repository for brain connectivity matrices	Brown, A. et al.
31 NeuroVault.org: a repository for sharing unthresholded statistical maps, parcellations, and atlases of the human brain	Gorgolewski, J. et al.
32 ANIMA: a data-sharing initiative for neuroimaging meta-analyses	Reid, T. et al.
33 A database of age-appropriate average MRI templates	Richards, J.E. et al.
34 The Cambridge MRI database for animal models of Huntington disease	Sawiak, J. et al.